Portuguese High School Student Math Performance Analysis

Vinay Venkatesh, Shaan Lehal, and Jimmy Zhang

Motivation

Academic performance in secondary school has long-term consequences for students' educational opportunities, career options, and overall life outcomes, yet student achievement is shaped by a complex mix of academic habits, family background, social environment, and personal behaviors. Understanding how these factors interact is not only of broad scientific interest, but also has practical implications: identifying which attributes most strongly relate to achievement can help educators and policymakers design more effective interventions, allocate resources, and support students who may be at risk of underperforming.

The UCI Student Performance dataset provides a rare opportunity to explore these relationships at an individual level. Unlike many educational datasets that include only grades and demographics, this dataset incorporates detailed information on students' family situations, study habits, lifestyle choices, and social relationships, along with multiple stages of academic performance (G1, G2, and the final grade G3).

Dataset Description

In this report we analyze the UC Irving Student Performance dataset. The data come from the UCI Machine Learning Repository's "Student Performance" dataset, which contains information collected from students in two Portuguese secondary schools. There are two separate datasets:

student-mat.csv — students enrolled in the Mathematics course student-por.csv — students enrolled in the Portuguese language course

For the purposes of this analysis, we will focus on the math dataset. This dataset contains 395 observations for Mathematics. Rows represent individual students, and columns represent attributes describing demographic characteristics, family background, study habits, lifestyle behaviors, and academic performance. The dataset contains 33 variables, which fall into the following groups:

Demographics and Family Background school (GP or MS) sex, age, address (urban/rural) famsize (GT3/LE3), Pstatus Medu, Fedu, Mjob, Fjob guardian, reason for school choice

Academic Engagement studytime, traveltime, failures schoolsup (school support), famsup, paid classes higher (wants higher education), activities

Lifestyle and Social Behavior goout, freetime, famrel Dalc (weekday alcohol use), Walc (weekend alcohol use) romantic relationship, health, absences

Academic Outcomes G1, G2, and G3 (final grade, the main outcome variable) Grades range from 0 to 20, with G3 serving as the final performance measure for each student in the respective subject.

Research Questions

In this analysis we will explore three research questions:

1) How does students' academic engagement affect their grades?

2) How do students' demographics and family background affect their grades?

3) How do students' lifestyle choices and behavioral patterns affect their grades?

If we answer these questions we can modularly figure out what factors are most predictive of student outcomes. Though each section of our report will focus primarily on variables that lie in one of these three groups, we will occasionally consider interactions between groups as they may reveal interesting patterns. We can then build a comprehensive model of student success.

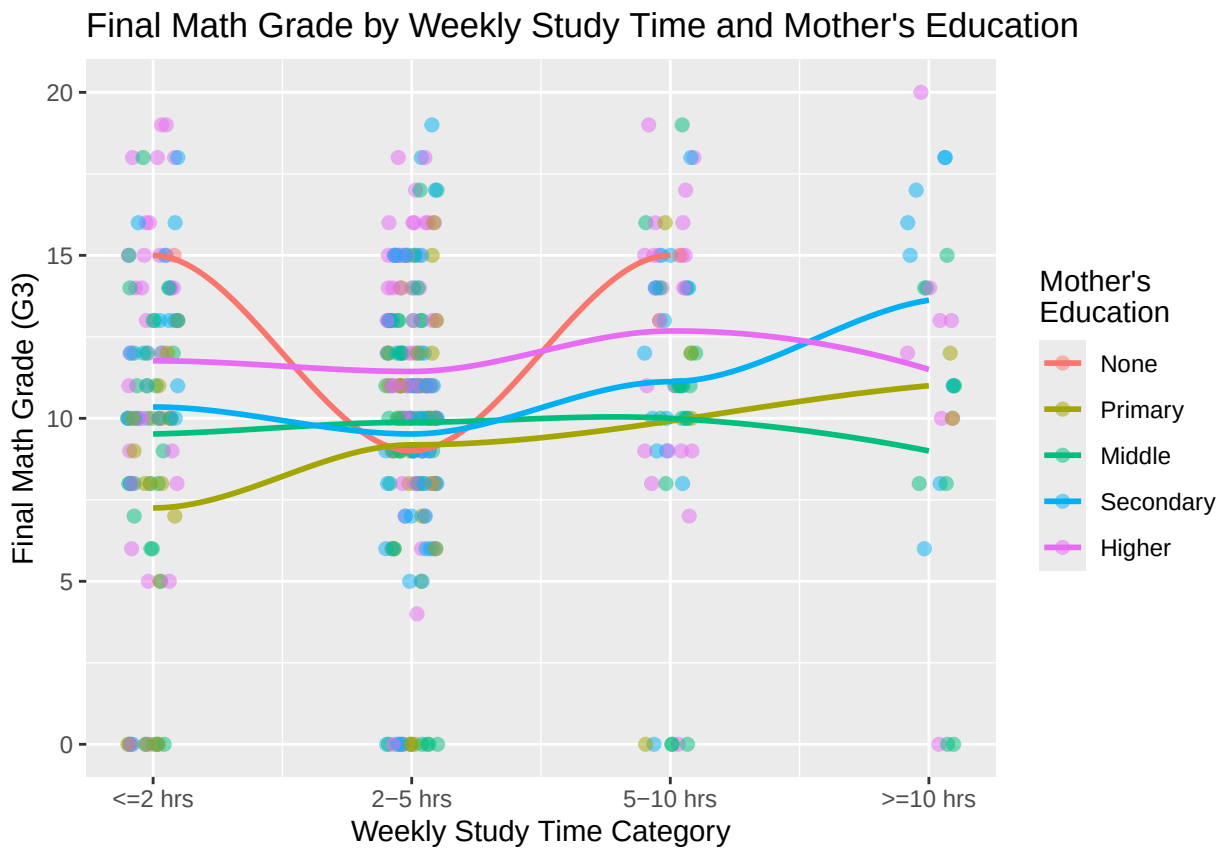
Graphs and Findings

1) How do student demographics and family background affect academic sucsess?

Intuitively, one of the most important factors that influence a student's educational outcome is their study habits. Investigating these behaviors helps illuminate whether students who invest more time in schoolwork actually see measurable benefits, and whether these benefits are uniform across different demographic groups. In particular, examining how study time interacts with parental education, how past failures shape current performance, and how grades evolve across the school year allows us to understand both the immediate and cumulative effects of study habits on achievement. Together, these analyses offer insight into where interventions may be most beneficial and which groups of students may be most at risk of falling behind.

Graph 1 – Study Time vs Final Grade, Colored by Mother's Education

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

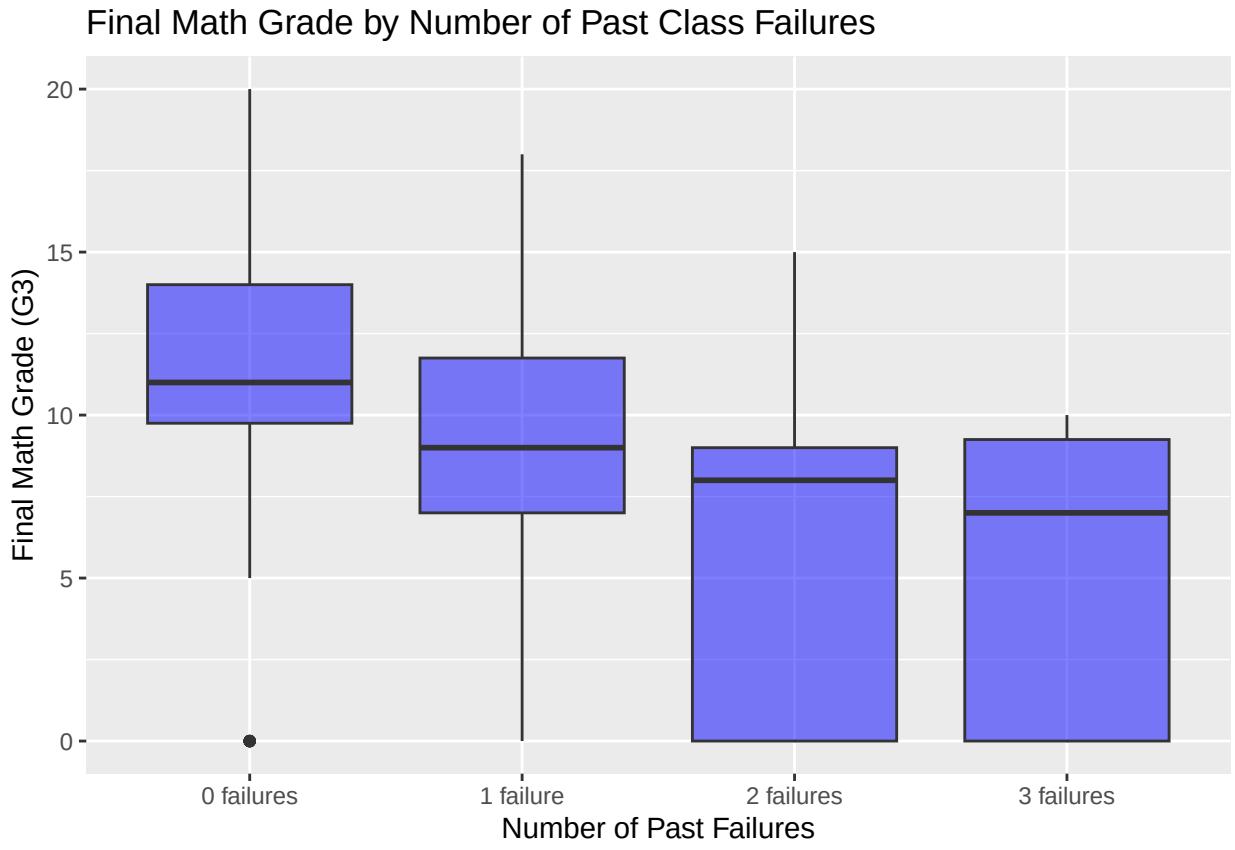


This visualization examines how weekly study time is associated with students' final math performance and how this relationship varies across different levels of maternal education. Study time is one of the most direct indicators of academic effort, but its effectiveness may depend on the level of support and educational capital available at home. By overlaying smooth trend lines for each category of mother's education, the graph allows us to simultaneously evaluate both the direct effect of studying and the contextual effect of family background. This makes the plot especially useful for assessing whether the benefits of increased study hours are distributed equally across students or shaped by parental education.

Overall, the graph shows that higher study time is modestly associated with improved final math grades, though the relationship is not strictly linear and contains considerable variability. More strikingly, the colored trend lines indicate that maternal education introduces substantial differences in performance: students whose mothers attained secondary or higher education consistently outperform peers with less educated mothers, even when they report similar study habits. This suggests that family background may amplify or condition the returns to studying, with students from more educated households achieving higher grades even at lower study-time levels. The results highlight that while studying matters, the academic environment at home plays an important supporting role.

Graph 2 – Past Failures vs Final Grade (Boxplot)

While study habits and family background offer important insight into why some students perform better than others, these factors represent only part of a student's academic story. Equally important is a student's prior record of achievement, which often reflects long-standing patterns in engagement, preparedness, and support. To understand how past academic difficulties influence current performance, we next examine the relationship between students' history of class failures and their final math grade.



The boxplot shows a striking decline in final grades as the number of past failures increases. Students with zero failures exhibit the highest median grades and the tightest distribution, while even one failure corresponds to a noticeable drop in performance. Those with multiple failures show markedly lower and

more variable scores, with very few students achieving high grades. The steep, stepwise decline in medians suggests that academic struggles accumulate and compound over time.

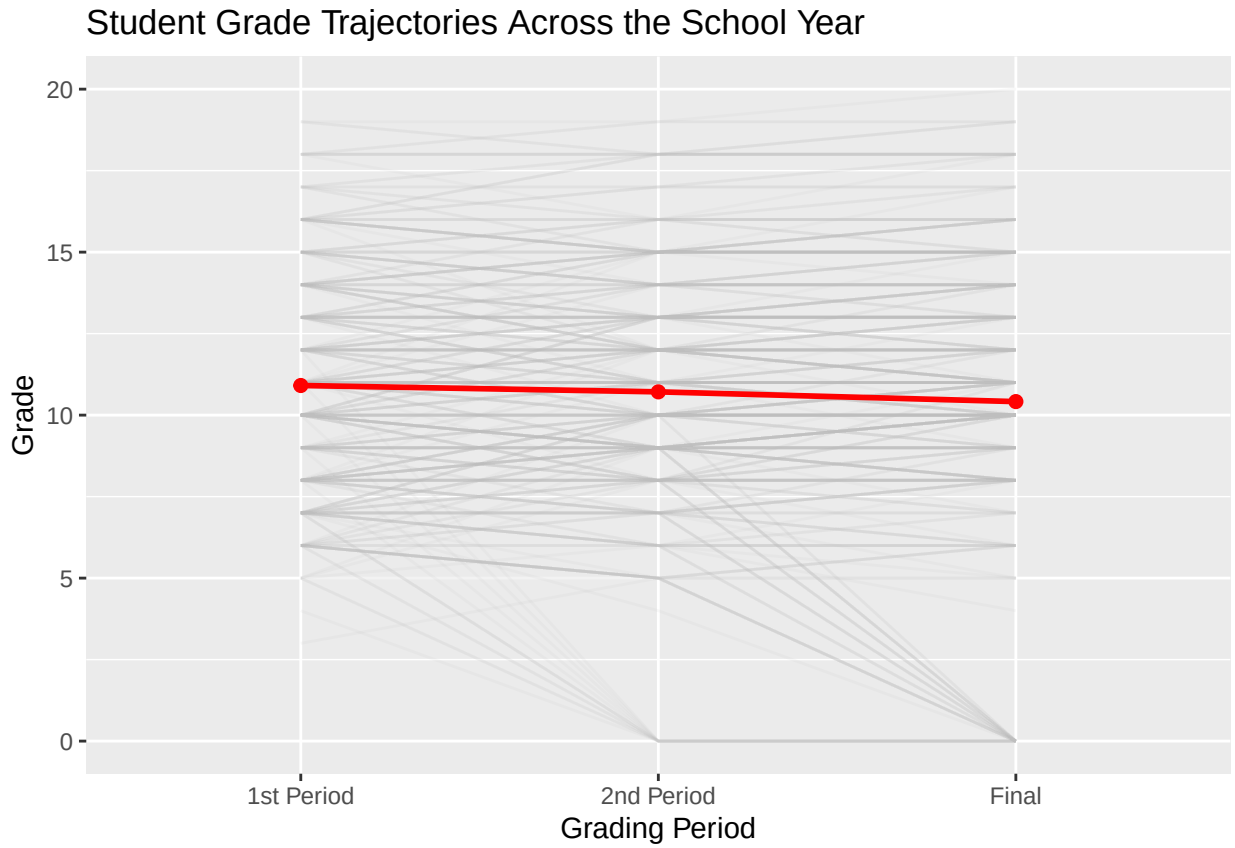
Because this visual pattern was so strong, we verified it statistically using a one-way ANOVA. The results confirm the visual impression:

```
##                   Df Sum Sq Mean Sq F value   Pr(>F)
## failures_factor    3   1137   379.0   20.78 1.64e-12 ***
## Residuals        391   7133    18.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This extremely small p-value indicates overwhelming evidence that mean final grades differ across failure categories. In other words, the number of past failures is not only visually associated with final performance—it is statistically one of the strongest predictors in the dataset. Students with even a single past failure are, on average, significantly behind their peers, and those with multiple failures show a pronounced achievement deficit.

Graph 3 – Grade Trajectories Across the Three Periods (G1 → G2 → G3)

Given the strong and statistically significant relationship between past failures and final grades, we next examine whether similar patterns emerge when we look within the school year itself. To do this, we analyze students' grade trajectories across the three grading periods.



This figure traces individual students' grade trajectories across the three main evaluation periods—G1, G2, and G3—providing insight into how academic performance evolves throughout the school year. By plotting each student's progression alongside the overall average trend, the graph highlights patterns of improvement, stability, or decline that reflect both study habits and broader engagement with the coursework. Studying

these trajectories helps us identify whether students tend to recover from early difficulties, maintain consistent effort, or drift downward as the material becomes more challenging.

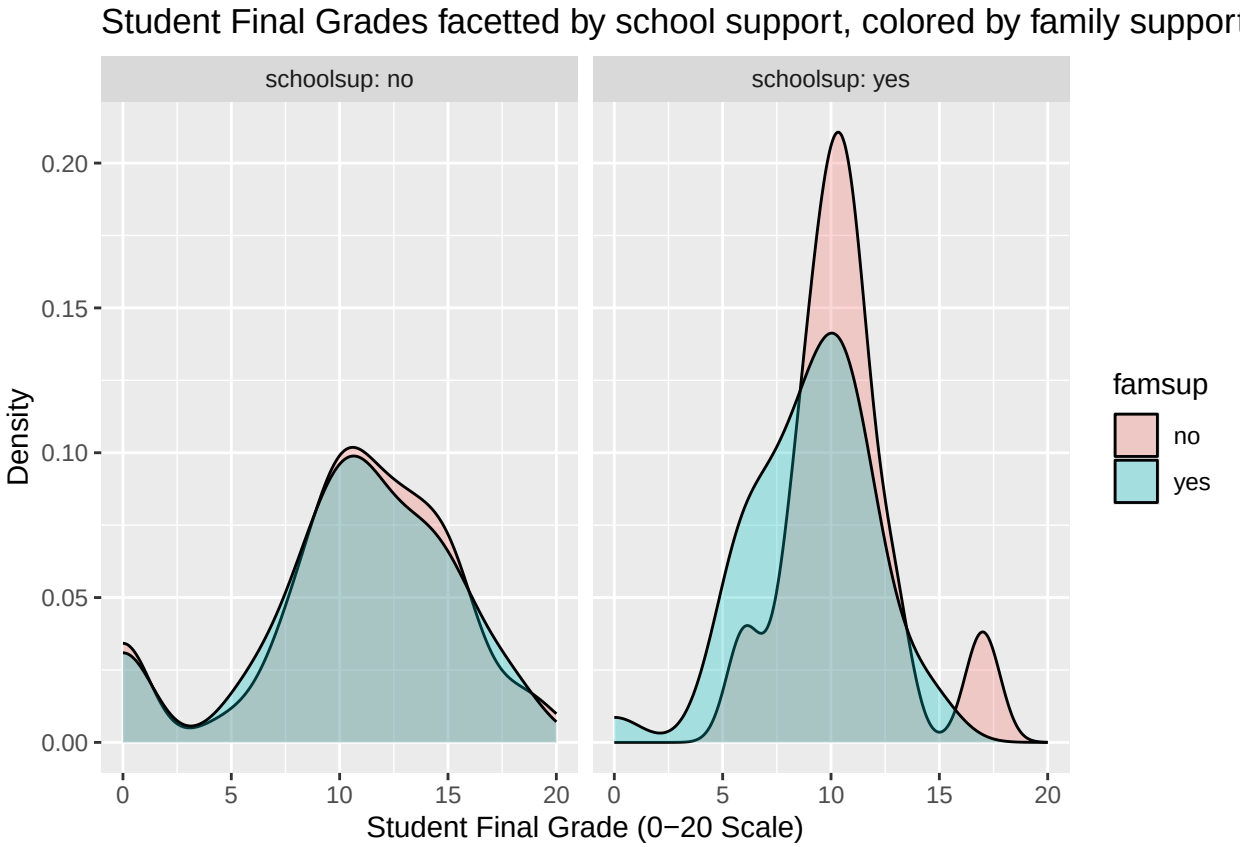
The individual trajectories reveal substantial heterogeneity, but many students exhibit stable or slightly declining performance over time. The red mean trend line shows a mild rise from the first to the second period, followed by a small decline in the final exam, indicating that early gains are not consistently carried through to the end of the year. A sizable fraction of students drop sharply at the final assessment, possibly reflecting increased difficulty or diminished study engagement late in the term. Combined with evidence from the previous graphs, these patterns suggest that consistent study habits and ongoing support are critical: once students begin to fall behind, substantial improvement becomes less common, and late-term declines can significantly impact final outcomes.

Research Question 2: How do students' demographics and family background affect their grades?

Academic achievement is shaped not only by the choices students make but also by the demographic and family environments in which they grow up. A student's access to educational support, the stability of their home environment, and the educational attainment of their parents all contribute to the opportunities and expectations that surround learning. Understanding how these background characteristics relate to academic performance is crucial for identifying structural sources of inequality and for determining which groups of students may benefit most from additional support. By examining both the institutional supports students receive and the resources available at home (such as parental education and family support), we can develop a more complete picture of the broader conditions that influence academic success. In this section, we analyze how these contextual factors correlate with final math grades to assess the extent to which family background and demographic characteristics contribute to student outcomes.

Graph 4 – Student Final Grades by school support and family support

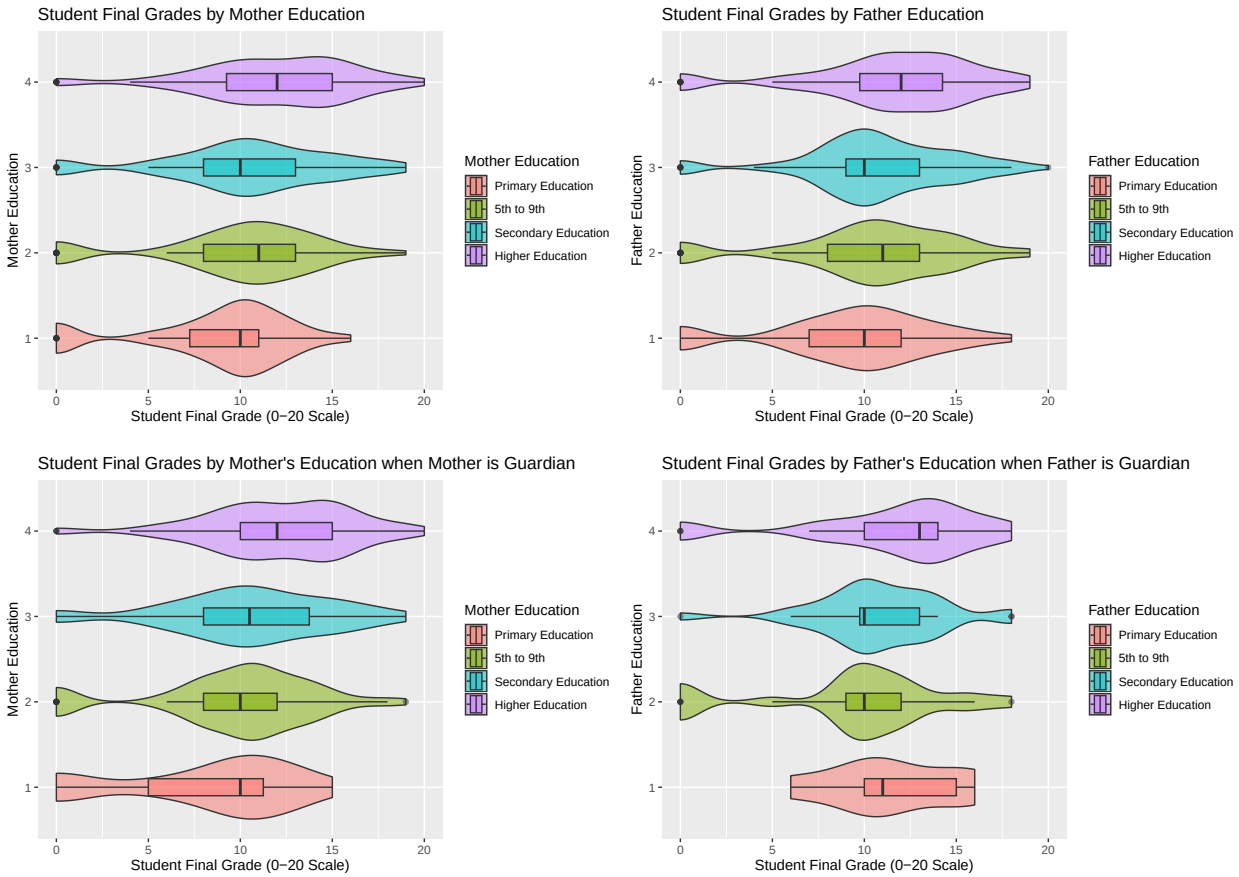
This density plot is designed to examine whether access to academic support—either from the school or from one's family—corresponds to differences in final math achievement. By faceting the plot by school support and coloring by family support, we can directly observe how these two key forms of assistance interact. This visualization is useful because it allows us to see not just average differences but the shape of grade distributions within each support configuration, revealing whether support systems benefit all students similarly, shift the entire distribution upward, or primarily help certain subgroups.



The density curves indicate that students who receive no school support have fairly similar grade distributions regardless of whether they receive family support, suggesting limited differentiation when institutional support is absent. However, among students who do receive school support, the presence of family support appears to shift the distribution of grades backwards, with supported students achieving higher densities around lower grade values. This pattern suggests that support systems may not be effectively working together: students receive less benefit when both their school and family environments provide assistance, rather than just school support. Another possible explanation is that the students that require additional school support already were not doing as well academically. These results emphasize that impacts on student performance may be unintuitive, with less support from family leading to higher grades.

Graph 5 – Violin Plot of Student Grades by Parental Education and Guardianship

Parental education is another one of the strongest predictors of children's academic outcomes across many educational systems. It reflects not just subject-specific knowledge but also broader forms of cultural and social capital, including familiarity with school expectations, the ability to assist with homework, and attitudes toward achievement and discipline. These violin plots visualize the distribution of student grades across different categories of mother's and father's education, providing insight into whether higher parental education corresponds to better student performance. The violin+boxplot combination allows us to examine both the central tendencies and the full distribution of grades within each educational category, revealing not just average differences but the spread and skewness of outcomes.



Here is an ANOVA test run on student final grade over mother education. The differences in means are statistically significant.

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## Medu          1    478   477.9   24.03 1.4e-06 ***
## Residuals   388   7716    19.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here is another ANOVA test run on student final grade over father education. The differences in means are also statistically significant.

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## Fedu          1    221  220.63   10.74 0.00115 **
## Residuals   388   7973   20.55
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Across both parents, the violin plots reveal a clear and consistent pattern: students with more highly educated parents tend to achieve higher final math grades. The distributions shift upward as parental education moves from primary schooling to higher education, with both the median and the upper tails showing notable improvement. Students whose mothers or fathers completed secondary or higher education display tighter, more favorable grade distributions, while those with parents who completed only primary schooling show lower medians and a greater concentration of low-performing students.

When the analysis is restricted to families where the mother or father is the primary guardian, the same upward trend persists. In these cases, parental education appears to have an even stronger association

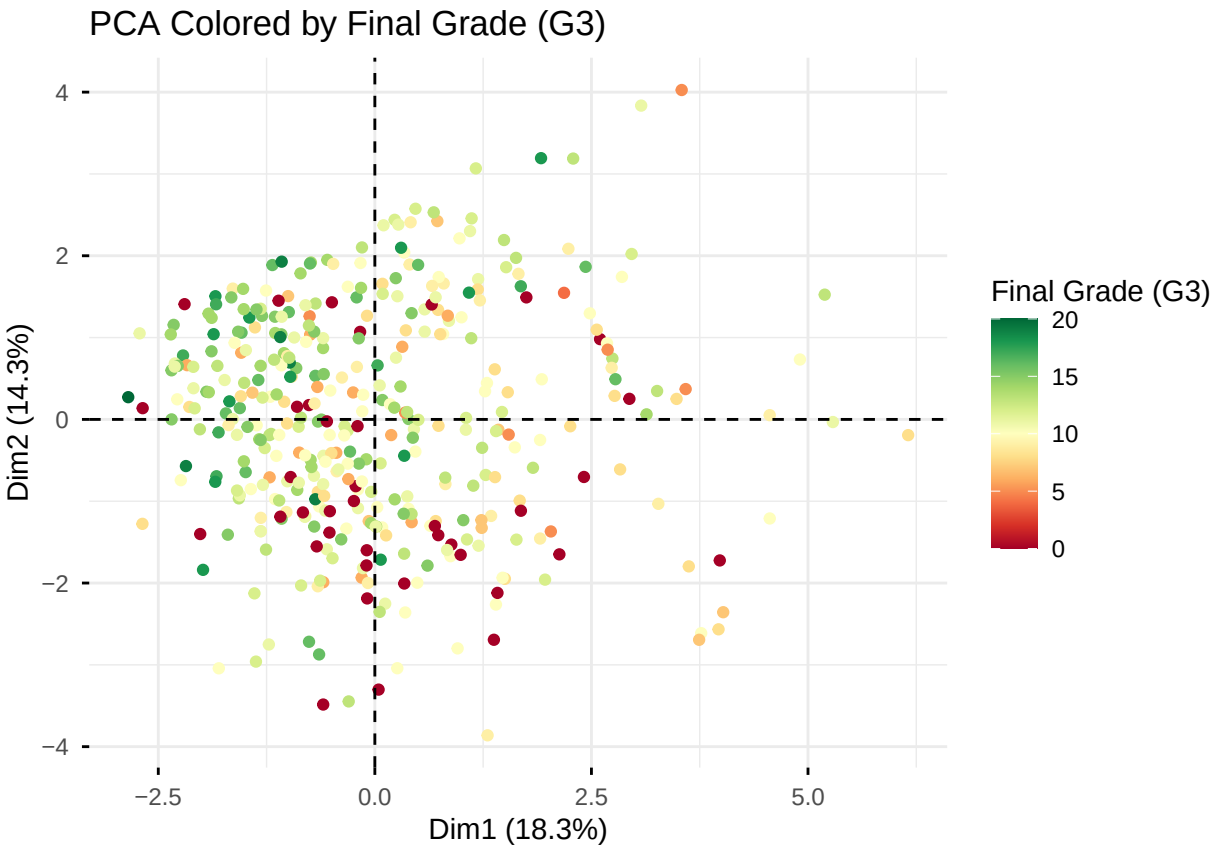
with student outcomes, suggesting that the influence of an educated parent may be amplified when that parent plays a central role in day-to-day supervision and support. Taken together, these results highlight the substantial role that family cultural and educational capital plays in shaping academic achievement, reinforcing the idea that interventions aimed at supporting lower-education households may have meaningful effects on student performance.

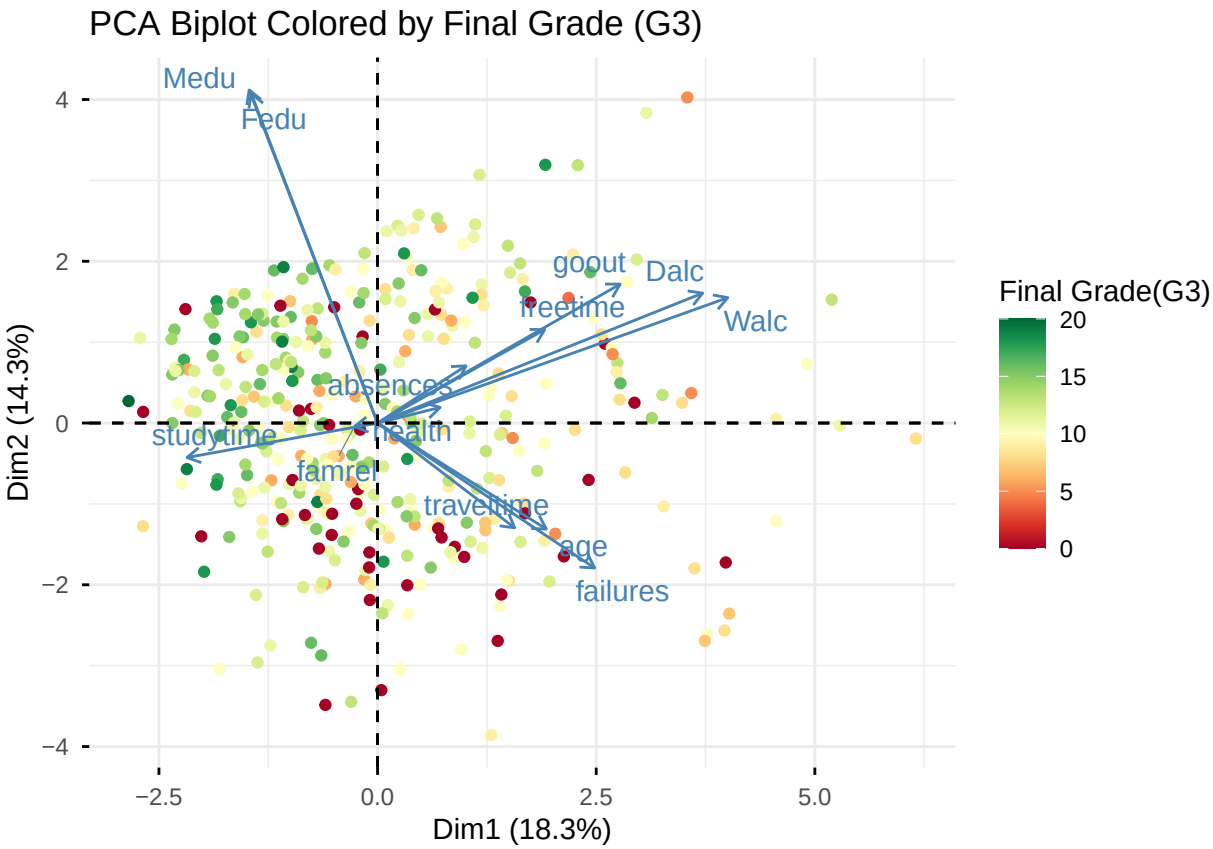
Research Question 3–How do students' lifestyle choices and behavioral patterns affect their grades?

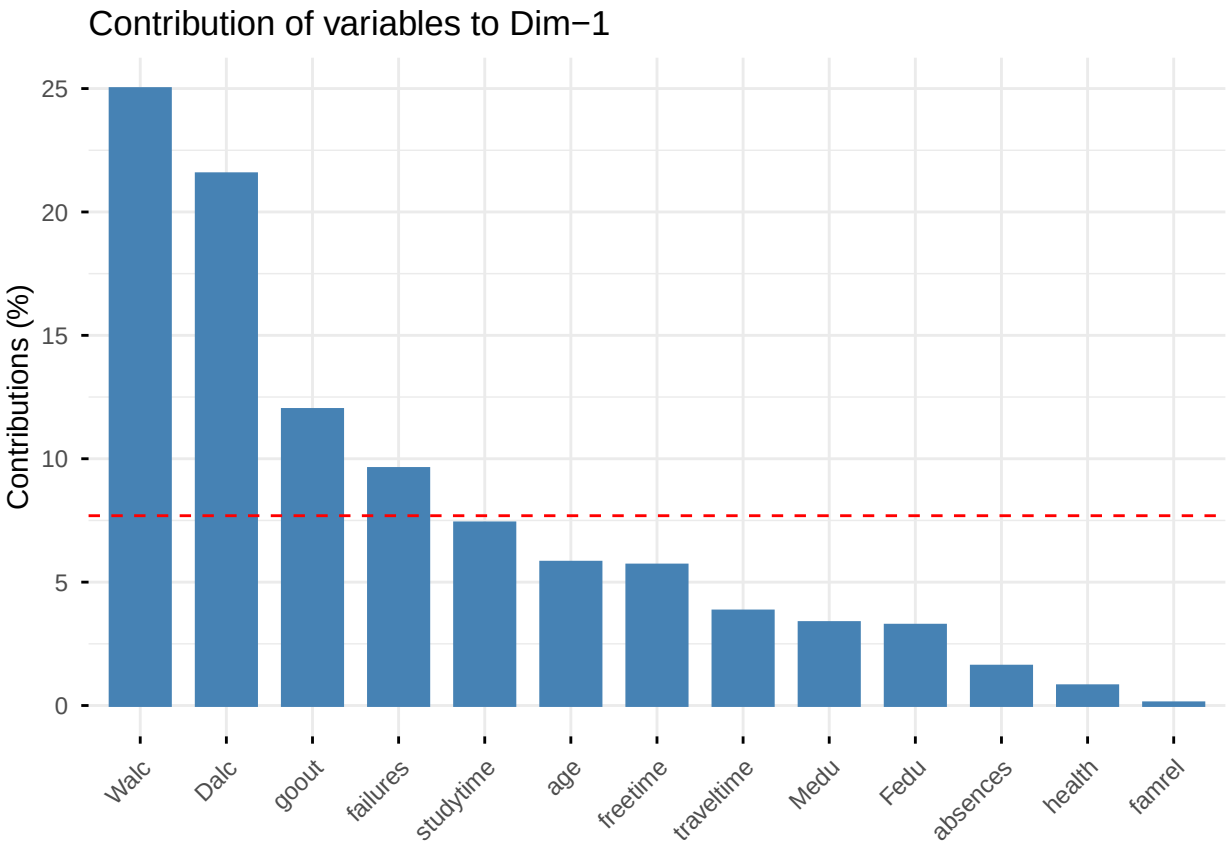
Beyond study habits and family background, students' everyday behavioral choices also play an important role in shaping academic outcomes. Behaviors such as alcohol consumption, social activity, and overall engagement can directly affect cognitive performance, sleep, motivation, and the ability to complete schoolwork. In adolescence, these behaviors often vary widely between students and can reflect differing priorities, peer groups, and external pressures. Understanding how these lifestyle choices relate to academic achievement is essential for identifying at-risk students and guiding interventions that promote healthier, more productive habits. To investigate this, we analyze a range of behavioral variables—including alcohol use, going out frequency, and broader lifestyle patterns—using both visualization and multivariate analysis.

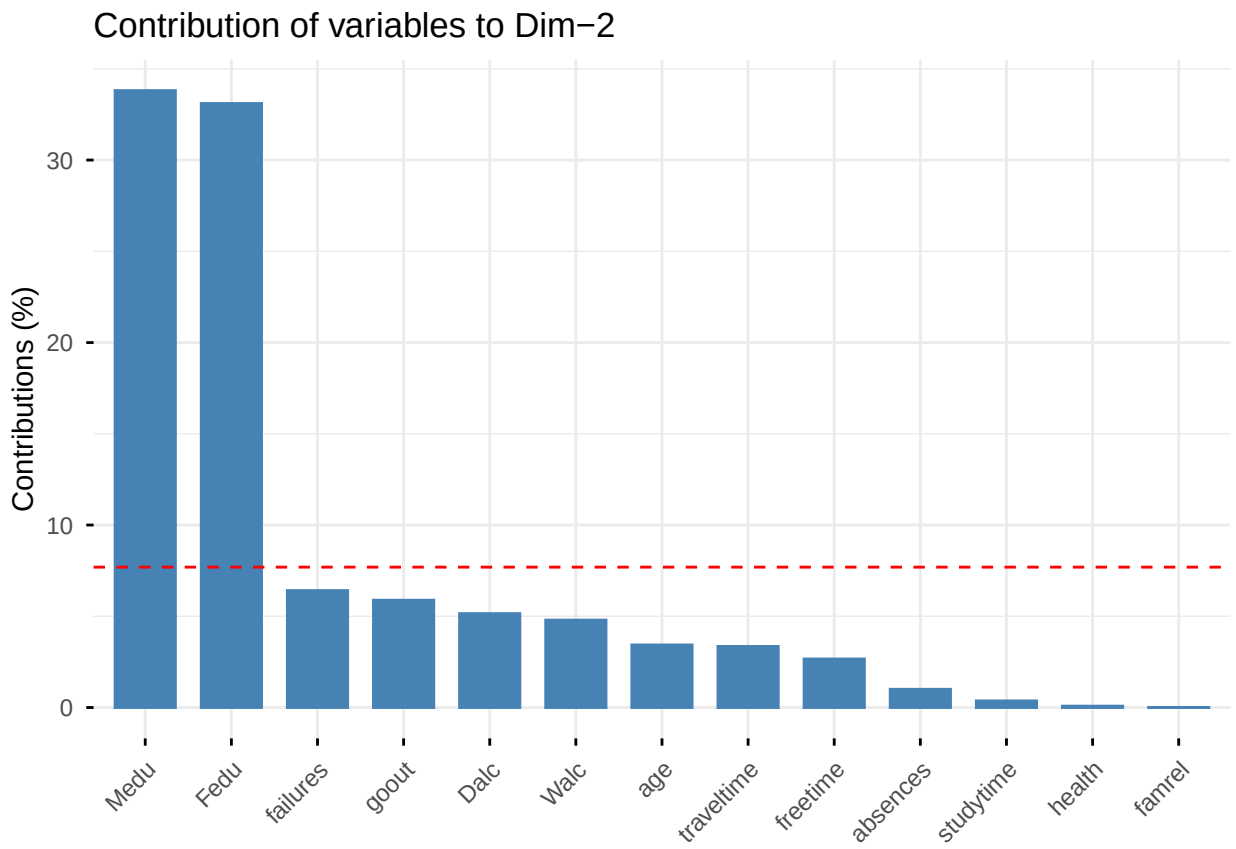
Graph 6–First two PCs of dataset (excluding grade variables) colored by final grade

The PCA (Principal Component Analysis) allows us to view students' behavioral and lifestyle profiles in a reduced two-dimensional space and examine how those patterns relate to final math grades. By projecting multiple behavioral variables—such as alcohol use, going out frequency, age, study time, and absences—into two principal components, the PCA helps us identify the dominant behavioral patterns in the data and see whether students with similar habits cluster together. Coloring the points by final grade (G3) reveals whether particular behavioral profiles are associated with stronger or weaker academic performance. The accompanying contribution plots show which variables drive each principal component, allowing us to interpret the behavioral dimensions more clearly.


PCA Colored by Final Grade (G3)
Dim2 (14.3%)
Dim1 (18.3%)
4
2
0
−2
−4
−2.5
0.0
2.5
5.0
Final Grade (G3)
20
15
10
5
0


PCA Biplot Colored by Final Grade (G3)
Medu
Fedu
goout
Dalc
freetime
Walc
absences
studytime
health
famrel
traveltime
age
failures
Dim2 (14.3%)
Dim1 (18.3%)
4
2
0
−2
−4
−2.5
0.0
2.5
5.0
Final Grade(G3)
20
15
10
5
0


Contribution of variables to Dim−1
Contributions (%)
25
20
15
10
5
0
Walc
Dalc
goout
failures
studytime
age
freetime
traveltime
Medu
Fedu
absences
health
famrel



The PCA scatterplot shows that student grades are spread widely across the behavioral space, with no sharp clusters corresponding to high or low performers. However, subtle patterns emerge: lower-performing students (red tones) tend to occupy regions associated with higher weekend and weekday alcohol consumption, whereas higher-performing students (green tones) appear more frequently in regions associated with moderate study habits, lower alcohol use, and more stable behavior.

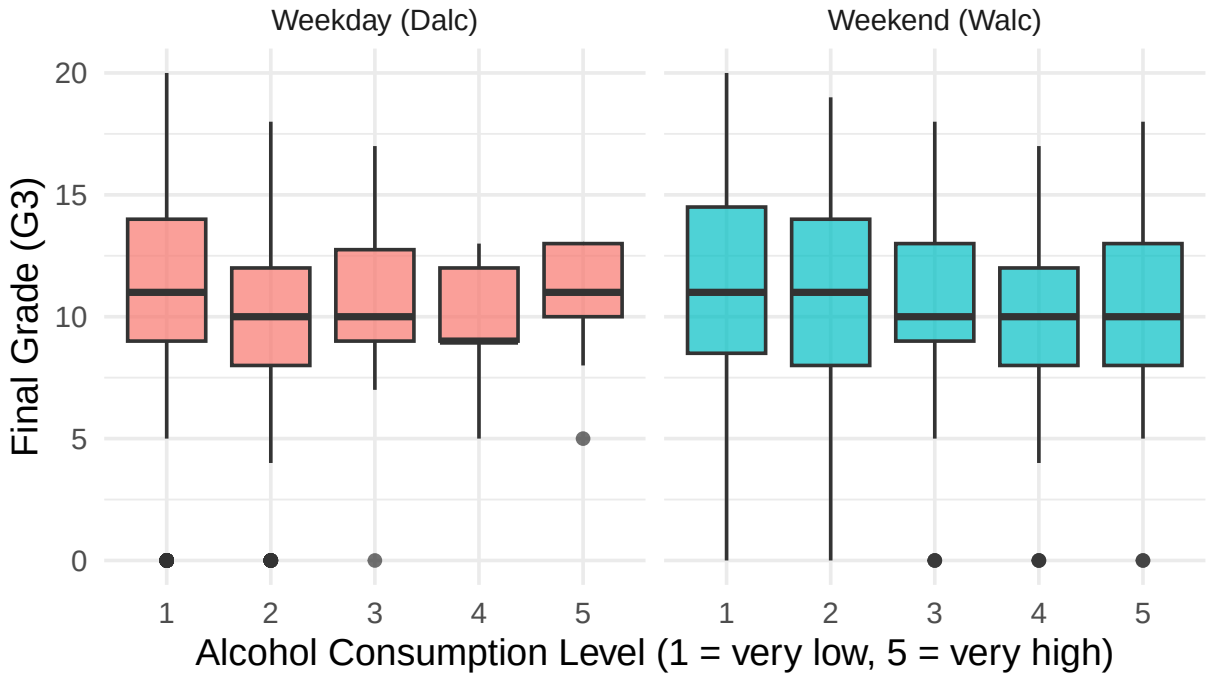
The contribution barplots along with the biplot of dim1 and dim2 reinforce this interpretation along.

Dim 1 is driven primarily by alcohol consumption (Walc, Dalc) and going out, suggesting that this component represents a "social activity and alcohol use" dimension.

Dim 2 is dominated by parental education, which is consistent with earlier findings but less central to behavioral choices.

Overall, the PCA reveals that behavioral decisions—especially alcohol consumption and social activity—form a key underlying dimension of variation among students. While these behaviors do not create sharply defined clusters of high vs. low achievers, they do correlate with performance patterns: heavier drinking and more outgoing lifestyles tend to align with lower grades.

Graph 7: Impact of Weekday vs Weekend Alcohol Consumption on Math Final Grades Because the PCA reveals that alcohol consumption—particularly weekend use—dominates the first principal component and is one of the strongest behavioral signals in the dataset, we next examine this variable more directly. While PCA provides an overall structural view of behavior, it does not show the precise relationship between alcohol use and final grades. To explore this connection in a more interpretable way, we move to boxplots that compare student grades across different levels of weekday and weekend drinking.



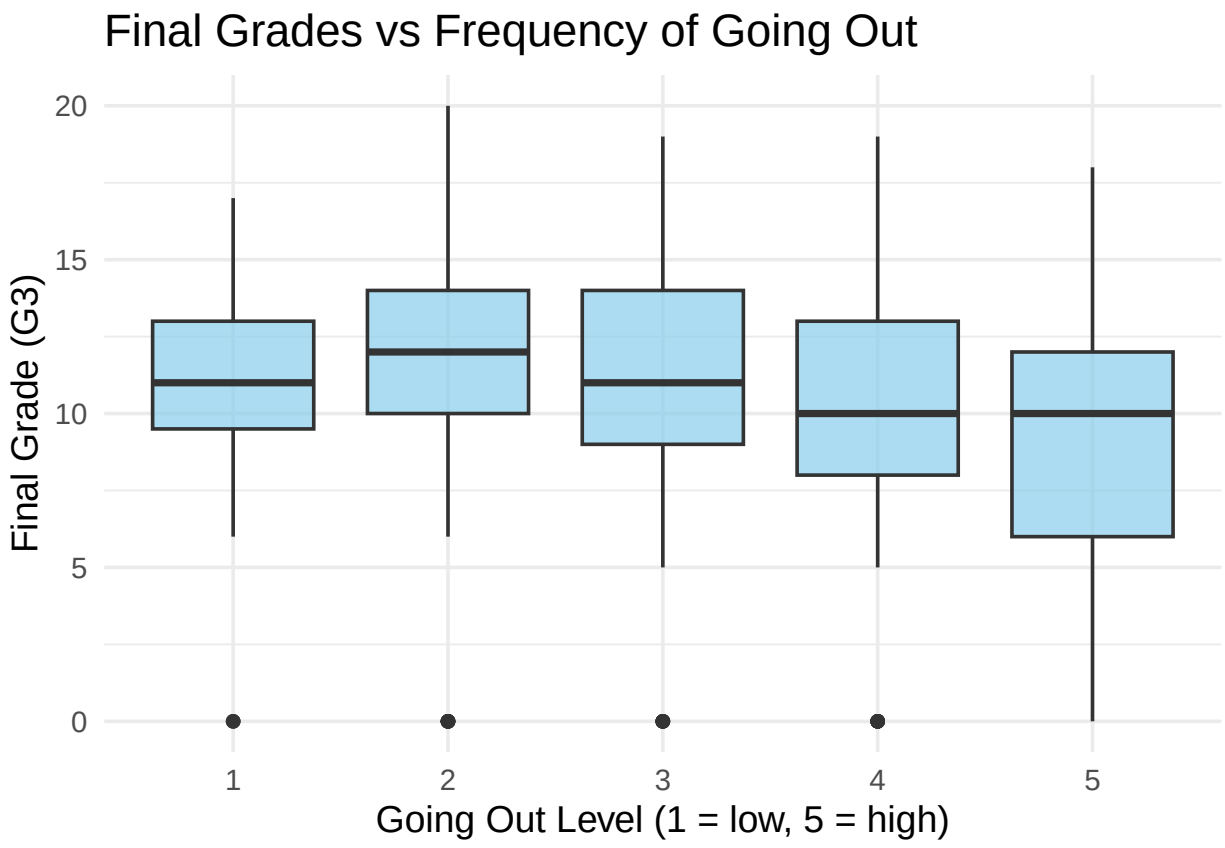

```
##
##  Spearman's rank correlation rho
##
## data:  df$Dalc and df$G3
## S = 11513871, p-value = 0.01618
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.1209445

##
##  Spearman's rank correlation rho
##
## data:  df$Walc and df$G3
## S = 11344535, p-value = 0.03797
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.1044586
```

Alcohol consumption is one of the most salient behavioral choices among older adolescents, with demonstrated effects on sleep, attention, and academic functioning. This pair of boxplots separately examines weekday (Dalc) and weekend (Walc) drinking to determine whether the timing of alcohol use matters for academic performance. By displaying grade distributions across increasing levels of consumption, the graph allows us to assess whether heavier drinking corresponds to lower achievement and whether weekday or weekend habits have stronger associations.

The boxplots reveal a downward trend with some variance: students who drink more heavily—either during the week tended to have a lower final grade with a slight increase in median for very high alcohol consumption. For weekend drinking, students with very low and low alcohol consumption have a clear higher final grade than higher alcohol consumption, though the whiskers don't decrease as much as weekday alcohol consumption. These visual patterns are supported by formal statistical tests. A Spearman rank correlation shows a significant negative association between weekday alcohol consumption and final math grade (rho = –0.121, p = 0.016), indicating that grades tend to decline as weekday drinking increases. Weekend alcohol use shows a similar but slightly weaker pattern (rho = –0.104, p = 0.038). These results confirm that the downward trends observed in the boxplots are not due to random variation: higher alcohol consumption is reliably linked to lower academic performance. The stronger correlation for weekday drinking is consistent with the idea that drinking during the school week interferes more directly with attendance, concentration, or study time. Combined with the PCA findings, these statistical results reinforce the broader conclusion that lifestyle choices—particularly alcohol consumption—play a meaningful role in shaping academic outcomes.

Graph 8: Final Grades Vs. Going Out Frequency

Since alcohol consumption is closely tied to broader social behavior, it is natural to ask whether other aspects of students' social lives show a similar relationship with academic performance. The PCA results highlight "going out with friends" as another major behavioral contributor, suggesting that social activity may represent a second dimension of lifestyle differences among students. Having examined alcohol use in detail, we now turn to going-out frequency to determine whether heavier social engagement parallels the patterns observed for drinking behavior.



The frequency with which students go out with friends reflects their social engagement, time spent away from schoolwork, and exposure to peer influences. This boxplot examines how students' final grades vary across increasing levels of going-out frequency, allowing us to assess whether social activity competes with

academic responsibilities or whether moderate socialization might be compatible with—or even supportive of—school performance.

The going-out plot reveals a modest but noticeable decline in grades among students who report the highest levels of social activity. Students who seldom or moderately go out (levels 1–3) show similar median grades, but performance begins to decline at levels 4 and 5, with wider score variability and more low-end outliers. This pattern suggests that moderate social behavior does not harm academic performance and may be part of a balanced lifestyle; however, frequent late-night or high-intensity social activity likely detracts from the time and focus needed for schoolwork. Together with the alcohol consumption results, this graph supports the broader conclusion that high-intensity social behaviors are associated with lower academic achievement.

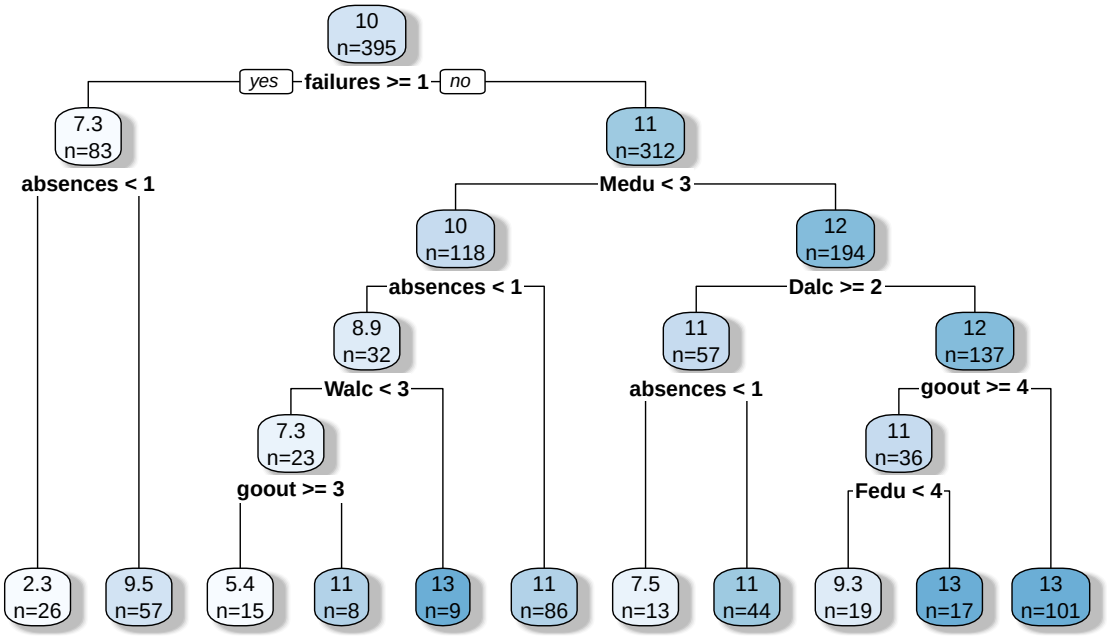
Decision Tree

Across this report, we examined three major dimensions of student performance: (1) study habits, (2) demographics and family background, and (3) behavioral decisions. Each section explored individual relationships between specific variables and final math grades (G3). However, academic performance is rarely driven by a single factor on its own. Instead, real-world outcomes arise from interactions—the way multiple characteristics combine and influence one another.

To synthesize the insights from these earlier analyses, we constructed a decision tree model using only the strongest predictors identified throughout the report: past failures, absences, parental education, study time, weekday and weekend alcohol consumption, and social behavior (going out). This final tree therefore serves as the culminating analysis of the report. It integrates the most meaningful predictors and reveals how they collectively shape a student's final grade.

```
##   absences   failures      goout       Walc       Dalc       Medu       Fedu
## 1236.79319 1042.74339  345.49492  252.96562  250.10497  232.86208  175.47546
##  studytime
##   11.30384
```

Decision Tree Predicting Final Math Grade (G3)



The decision tree reveals a clear hierarchy of the strongest predictors of student performance. The first and most influential split occurs on past class failures: students with one or more failures form an immediate low-performing group, confirming earlier analyses showing that prior academic struggles are one of the steepest predictors of reduced final grades. Within this at-risk group, the tree shows that absences, weekend alcohol use, and going-out frequency further separate low and moderately performing students, mirroring patterns seen in the behavioral analysis where heavier drinking and social activity generally corresponded to lower achievement.

For students with no prior failures, the structure of the tree changes, and differences are explained more by family background and moderate behaviors. Here, mother's education becomes the next major decision point: higher maternal education consistently leads to higher predicted grades, echoing the violin plot findings from RQ2 that parental education provides a substantial academic advantage. Among students whose mothers have less education, weekday alcohol consumption, absences, and going-out behavior again influence outcomes, underscoring that behavioral choices can diminish the academic benefits that come from having no previous failures.

Overall, the decision tree integrates all three research areas—study habits, family background, and behavior—into a coherent predictive model. It highlights that strong academic performance emerges from a combination of academic history (no prior failures), stable engagement (low absences), supportive background (higher parental education), and responsible behavioral choices (lower drinking and moderate social activity). Conversely, the lowest-performing profiles cluster where academic risk, poor attendance, and risky behavior overlap. This final model brings together the report's earlier findings and demonstrates how these factors interact to shape real differences in student outcomes.

Conclusions

This project examined three central dimensions of student academic performance: study habits, family background, and behavioral choices. Across all analyses, several consistent themes emerged. First, indicators of academic engagement—especially past failures and absences—proved to be among the strongest predictors of final math grades. Students who begin the year with a history of academic difficulty or inconsistent attendance face considerably steeper challenges, as shown through both the boxplots and the decision tree. Study time itself showed modest gains, but its effects were overshadowed by engagement factors, confirming that studying alone cannot fully compensate for foundational academic struggles.

Second, family background, particularly parental education, emerged as a powerful and persistent influence. Students from households where mothers or fathers had higher levels of education consistently achieved stronger outcomes, even when controlling for behavioral factors. The violin plots and decision tree together highlight how parental education not only raises average grades but also narrows the spread of performance, suggesting more stable academic environments.

Third, behavioral decisions, including alcohol use and social activity, displayed meaningful—even if more moderate—relationships with performance. Heavier drinking, especially on weekends, and high going-out frequency were associated with reduced final grades. The PCA results reinforced these trends, identifying alcohol consumption and social behavior as major contributors to overall variation in student characteristics. Although these behaviors were not as influential as failures or parental education, they consistently appeared in lower-performing branches of the decision tree, demonstrating that lifestyle choices can magnify or mitigate existing academic risks.

Taken together, the decision tree provided a unified model synthesizing all findings. It revealed how engagement, family background, and behavior interact to create distinct academic trajectories. The strongest-performing students tend to combine no prior failures, consistent attendance, and supportive family environments, while students who face challenges in multiple domains show substantially lower predicted outcomes. These conclusions are well-supported by the analyses and address the research questions fully.

Further Questions

Although this report answers the core research questions, several important extensions remain outside the scope of the present analysis. One direction for future work involves exploring causal relationships rather than associations. Our methods—correlations, ANOVA, decision trees, and PCA—are descriptive and predictive, but they cannot establish whether behaviors such as alcohol use or study time cause changes in academic performance. More advanced techniques such as or instrumental variable analysis would be required to draw causal conclusions.

Another promising direction requires additional data. Many potentially influential factors—such as student motivation, teacher quality, mental health, peer networks, or socioeconomic status—are absent from the dataset. Including such variables could dramatically strengthen predictive accuracy and provide a more holistic picture of the forces shaping academic outcomes. Moreover, collecting similar datasets across multiple schools or countries would allow for cross-population comparisons and improve the generalizability of the conclusions.

Finally, future analyses could explore more sophisticated machine learning models, such as random forests, gradient boosting, or regularized regression. While the decision tree is transparent and interpretable, more complex models could capture nonlinear interactions and improve prediction performance. However, these models require deeper statistical training and were not emphasized in the current course, making them appropriate targets for future work.

Overall, future research should combine richer datasets with more advanced statistical methods to build on the foundation established here. Such extensions would help answer the nuanced questions that remain about how students' environments, habits, and decisions collectively shape their academic success.